x00111602

IT TALLAGHT  Conor Griffin

CA2: Detect patients that have asd

report by: conor griffin x00111602

Contents

[1. Section 1 2](#_Toc511739350)

[1. Analyse the dataset 2](#_Toc511739351)

[2. Prepare a number of views (formats) of the dataset 3](#_Toc511739352)

[Original Dataset 3](#_Toc511739353)

[Normalised View 3](#_Toc511739354)

[Standardised View 3](#_Toc511739355)

[Missing View 4](#_Toc511739356)

[3. Attribute Selection (Feature selection) 5](#_Toc511739357)

[4. Use kNearestNeighbor (IBK) Classifier on the dataset 8](#_Toc511739358)

[5. Two Machine Learning Algorithms for Evaluation 10](#_Toc511739359)

[Rules.DecisionTable 10](#_Toc511739360)

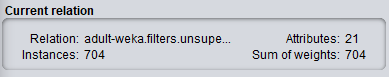
[Trees.SimpleChart 10](#_Toc511739361)

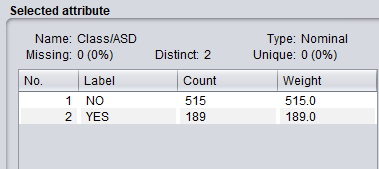
[6. Carry out an initial evaluation on Machine Learning algorithms 11](#_Toc511739362)

# Section 1

## Analyse the dataset

From looking at the dataset in Weka explorer there are a few things to take notice of. Looking at the current relation section in weka explorer it tells us that there are 704 rows of data and 21 columns/attributes in the dataset. For the attributes question 1-10 with a scale value of nominal type, each attribute has no missing values and no unique values, also each attribute has 2 distinct values. Another thing to take note of is that there are only 2 numeric attributes in the dataset, this can be a negative aspect in a dataset for reaching conclusions as it is good to have numeric attributes because they give a machine learning algorithm constant and final values to use. For each of the numeric attributes they have almost no missing values, age has just 2 missing values. Also the result numeric attribute has a minimum 0, maximum 10, Mean 4.875, Standard deviation 2501 this gives an idea of the spread of data. The attribute relation has a high amount of missing values 95 (13%) this will be interesting to see how algorithms interpret this amount of missing data and will this ratio of missing data mean I will have to remove or impute the values. The class attribute determines if an adult is autistic, “Yes”, or not, “No”, there is a higher count of patients in the No category 515 and 189 Yes. I would say that this is an accurate dataset in the sense you would expect to see a higher count in the no category based off the 10 questions.

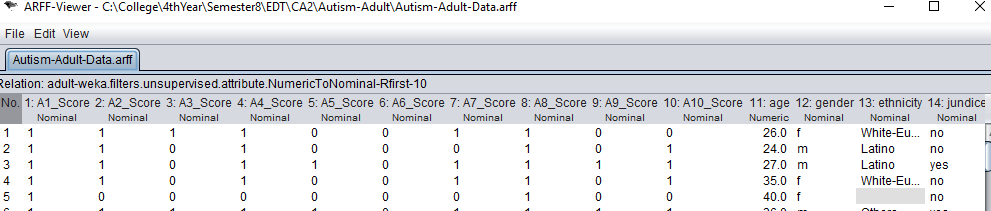




## Prepare a number of views (formats) of the dataset

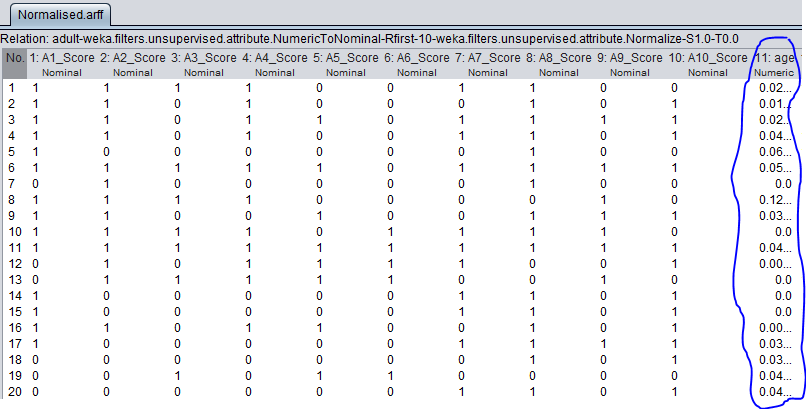
### Original Dataset

I loaded the dataset into Weka Arff viewer.



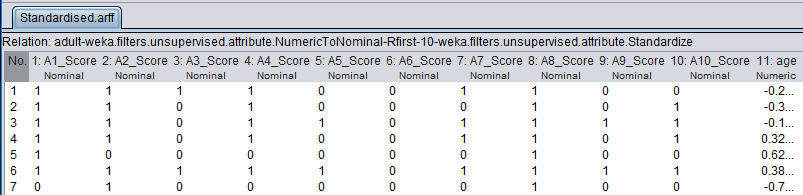
### Normalised View

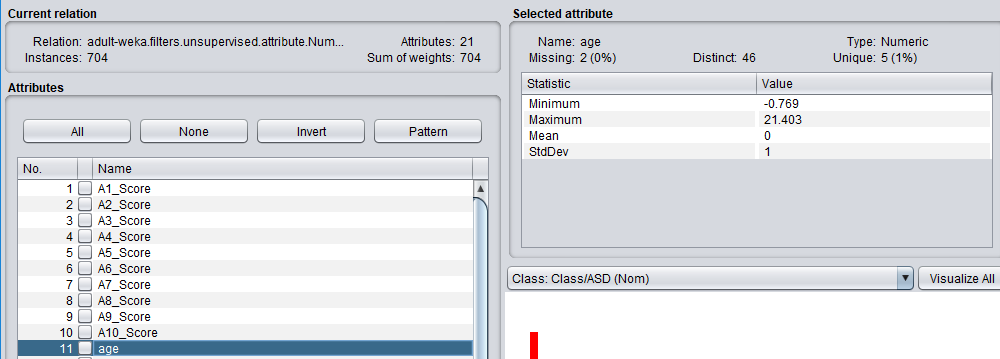
To normalise the view I first created a duplicate of the original dataset, then loaded it into weka and began the process in weka explorer. Data normalisation is the process of rescaling one or more numeric attributes to the range of 0 to 1. This is a good technique to use when you do not know the distribution of your data or when it is not a bell curved distribution. In weka explorer I chose the normalize filter and applied it to the autism dataset, renamed to normalised. From the final view you can see that the age numeric attribute has been normalised from ranges 0 to 1.



### Standardised View

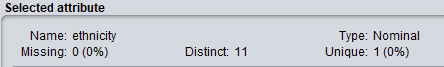
To standardise the view I began by creating a copy of original dataset called standardised.arff and loaded it into weka explorer. Data standardisation is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1. I standardised each of the attributes in weka by choosing the standardize filter and applied it accordingly. From the view you can see that the numeric attributes have been standardised, also in explorer you can see for example the age attribute has a mean of 0 and standard deviation 1.



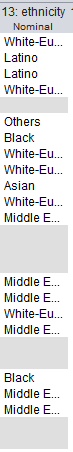


### Missing View

To impute the average into attributes with missing data I first loaded the dataset into weka and created a missing.arrf file. I then loaded it into weka explorer and used the ReplaceMissingValues filter. The results replaced all missing values with the average value. For example you can see below the attribute ethnicity’s values were imputed with the average ethnicity.



Normal view Missing View

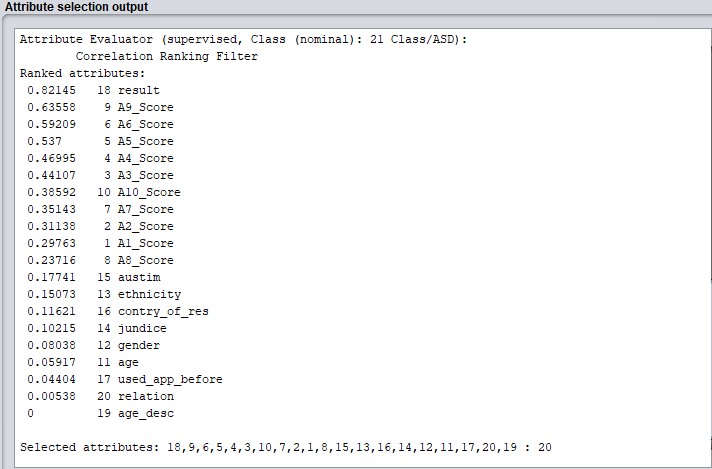
 

## Attribute Selection (Feature selection)

For attribute selection I divided the process into two parts: Attribute evaluator, this is used to evaluate the attribute in the context of the class variable. The Search Method, the technique to try or navigate combinations of attributes in the dataset in order to get on a short list of chosen features.

The first attribute selection method I used to determine which attributes to choose is the correlation attribute evaluation technique. I used this method with a Ranker search method which evaluates each attribute and lists the results in rank order. I configured both the attribute evaluator and search method to work with each other in Weka. I ran the algorithm on the dataset and go the following results.

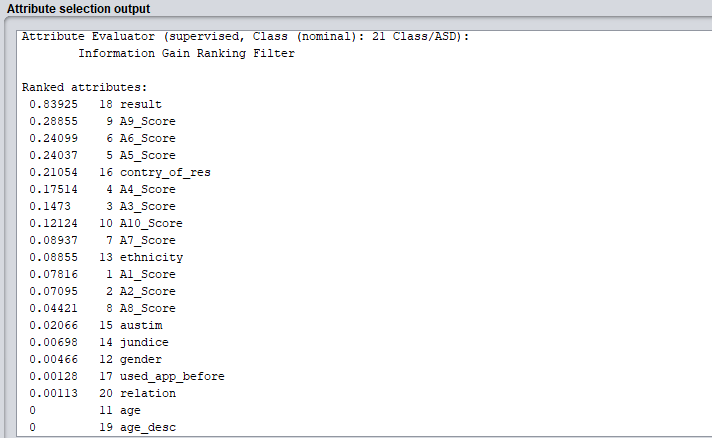




The technique to use here is to take a look at the correlation in attributes. Correlation is calculated for each variable predictor variable and the process to follow is to select only those attributes that have a moderate to high positive or negative correlation (close to -1 or 1) and drop those attributes with a low correlation (values close to 0). The use of the ranker search method also displays the attributes in a ranked order. Looking at the output you can see that the predictor attribute result has a really high correlation (0.82145) with the output class, also A9\_Score, A6\_score, A5\_Score, A4\_Score and A3\_Score have a high correlation with the class/ASD variable. If we set the cut-off point at 0.4 for relevant attributes then the remaining 14 attributes could possibly be removed, but this seems like a lot of data to remove so I would either set the cut-off lower or use other techniques to make a final decision on what attributes to remove.

The second method I used for attribute selection is the Information gain technique, which is used to calculate the info gain for each attribute for the output variable ASD. Entry values from 0 (no information) to 1 (maximum information). Those attributes that contribute more information will have a higher information gain value and can be selected whereas those that do not add much information will have a lower score and can be removed. In Weka I use the InfoGainAttributeEval in attribute evaluator and again use the Ranker search method. The following output displays the results.

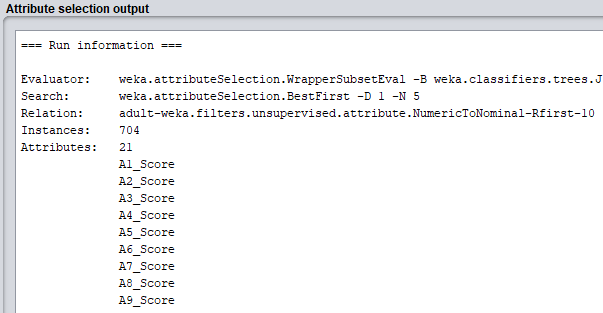


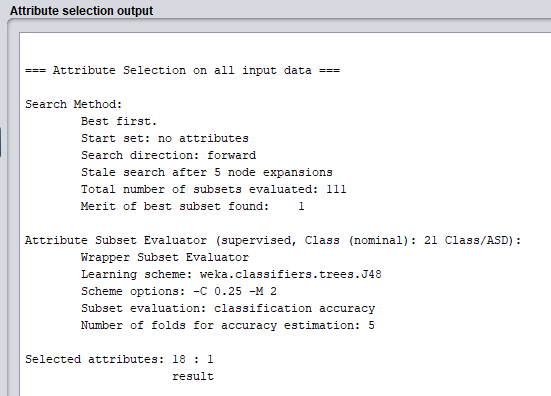


From the output you can see that the top ranked attribute is Result with a high information gain of 0.83925, most of the other attributes have a low info gain, but if we set the cut-off point at 0.2 then we could keep other attributes like A9, A6, A5\_Score and contry\_of\_res and the rest could be removed from the dataset.

The third selection method I used to determine what attributes to use in the modelling stage is the learner based feature selection. This is a powerful learning algorithm and is used to gather subsets of attributes, the subset with the best results and performance is taken as the selected subset. The feature selection method used is the WrapperSubsetEval technique and uses a BestFirst search method as it uses less compute time. I modified the configuration to instead of ZeroR, use J48 in trees, this gives a better preferred subset in the output.

The run information for this algorithm tells us that it ran the subset method with a best first search and used 704 instances over 21 attributes, it found 111 subsets and the merit of the best subset found was 1. The final selected attributes was attribute 18 result.

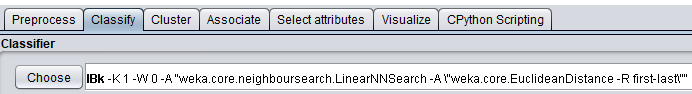




Looking back over the three techniques, you can see that a few of the attributes overlapped into each of the methods used. The results attribute would be the strongest and best attribute to keep in the dataset based on the three techniques used as it was number 1 in the output of each algorithm. Most of the other attributes didn’t perform well in each of the feature selection methods but this doesn’t necessarily mean you can get rid of all them, I would say it best to keep attributes that performed well in the first two techniques like A9\_Score, A6\_score, A5\_Score, A4\_Score and A3\_Score as well as contry\_of\_res attributes. I would add each of these in the modelling stage.

## Use kNearestNeighbor (IBK) Classifier on the dataset

Nearest neighbour: to classify a new instance, search training set for one that’s most like it. Searches the training set for the one that most like the new instance. We have our class variable with a yes or no answer, we now have the unknown instance, and the nearest neighbour method produces the new instance based on previous data most similar to it. It uses a similarity function by measuring the Euclidean distance, Manhattan distance, normalize the attributes to be between 0 and 1 and this gives us the kNearest Neighbor. KNearest protects us from noisy data, we choose the k nearest neighbors and it chooses the majority class among the neighbours.



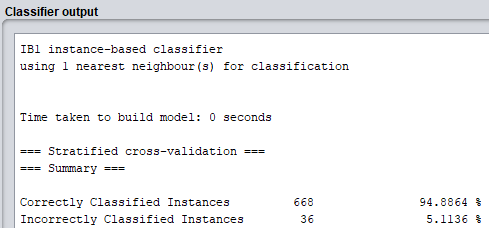
First, I found the baseline accuracy for the dataset, I ran the rules->ZeroR algorithm on the dataset, and this gave an output of 73.15%. I was able to base my accuracy of kNearest results off of the ZeroR result of 73%.

ZeroR

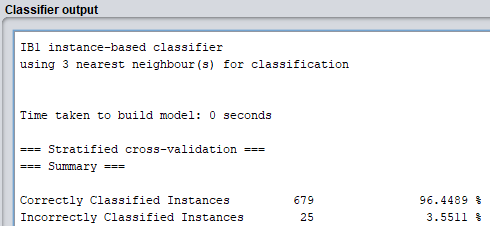


If any of the values for k have a greater accuracy than baseline performance then the more optimal that solution is. For k=1 the number of correctly classified instances was 668, incorrect classified instances was 36, with a 94% accuracy, k=3 the number of correctly classified instances was 679, incorrect instances was 25 with a 96% accuracy, k=5 the number of correctly classified instances was 675, incorrect instances was 29, with an accuracy of 95%, k=7 the number of correctly classified instances was 685, incorrect instances was 19, with an accuracy of 97% and k=15 the number of correctly classified instances 690, incorrect instances was 14, with an accuracy of 98%.

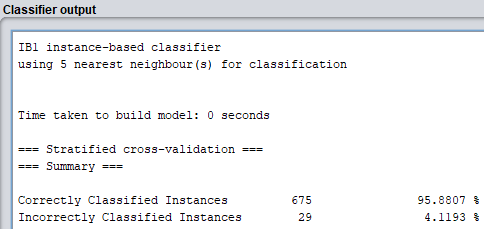
K=1



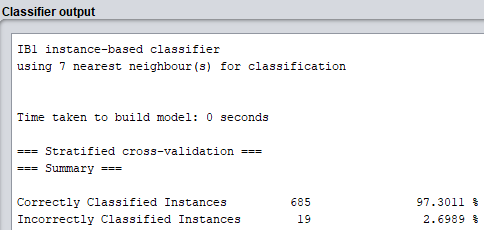
K=3



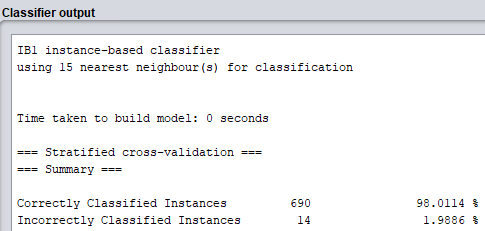
K=5



K=7



K=15



The optimal value of k I would say from looking at the results is 15 as it gave the most accurate % classified instances and more instances were produced (690). It seems that from looking at the results the greater the value for K then the better and more accurate the output but this is not the case, I also tested for greater values of k such as 20 and 50, the accuracy of these models were less than the model of k=15. Using 20 nearest neighbours for classification produced 689 correct instances and a 97% accuracy, using 50, produced 686 correct instances and 97% accuracy, both models less accurate than using 15 nearest neighbours for classification.

## Two Machine Learning Algorithms for Evaluation

### Rules.DecisionTable

<https://en.wikipedia.org/wiki/Decision_table>

<https://pdfs.semanticscholar.org/ad13/187dc62e8dd39e767258c7e70767733d54e5.pdf>

This machine learning algorithm uses a class for building and using a simple decision table majority classifier. Decision tables are one of the simplest hypotheses used in supervised machine learning algorithms. An algorithm inducing decision tables can sometimes outperform state of the art algorithms such as C4.5. It is one of many supervised machine learning algorithms that seek a hypothesis that will correctly predict the class of future unlabelled instances. In the case of the Autism-Adult dataset the decision table algorithm can be used to identify some of the unknown instances class identifier.

Decision tables give a visual representation which specifies which actions to perform depending on given conditions. The information expressed in decision tables could also be used in a programming language as a series of if-then-else and switch-case statements. Each decision in a decision table algorithm corresponds to a variable, relation or predicate whose possible values are listed among the condition alternatives. Each action is a procedure or operation to perform, and the entries specify whether (or in what order) the action is to be performed for the set of condition alternatives the entry corresponds to.

### Trees.SimpleChart

<https://pdfs.semanticscholar.org/26d6/73f140807942313545489b38241c1f0401d0.pdf>

Simple chart is a classification tree algorithm used in weka. It is a data mining algorithm that creates a step by step guide for how to determine the output for a new data instance. The tree it creates is exactly that: a tree whereby each node in the tree represents a spot where a decision must be made based on the input, and you move to the next node and the next until you reach a leaf that tells you the predicted output.

Simple cart method is CART analysis which stands for Classification and regression trees. It is used for data exploration and predication. It is a classification technique that generates the binary decision tree. Output is a binary tree, it generates only two children. Entropy is used to choose the best splitting attribute. Simple Cart handles the missing data by ignoring that record. This algorithm is best for the training data. CART decision tree is a learning technique, which gives the results as either classification or regression trees, depending on categorical or numeric data in the dataset. It is a greedy algorithm in that it chooses the locally best discriminatory feature at each stage in the process. This is suboptimal but a full search for fully optimized search would be computationally very expensive. In the CART decision tree the dataset is split into two subgroups that are the most different with respect to the outcome. This procedure Is continued on each subgroup until some minimum subgroup size is reached.

## Carry out an initial evaluation on Machine Learning algorithms

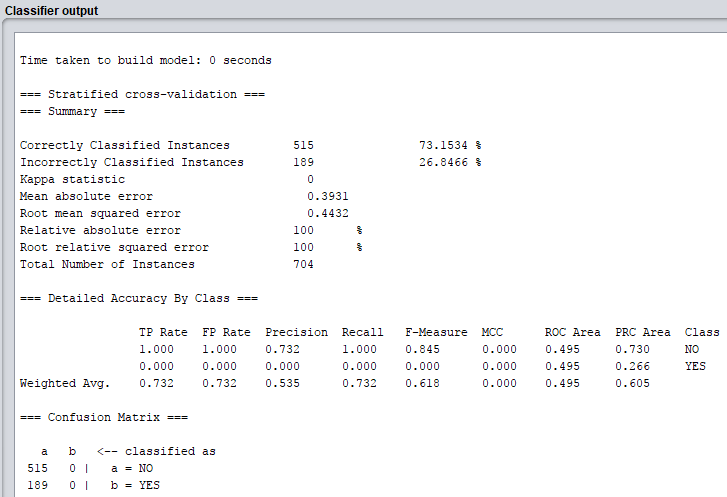
(result predictor removed)

<https://stackoverflow.com/questions/2903933/how-to-interpret-weka-classification>

### Rules.ZeroR

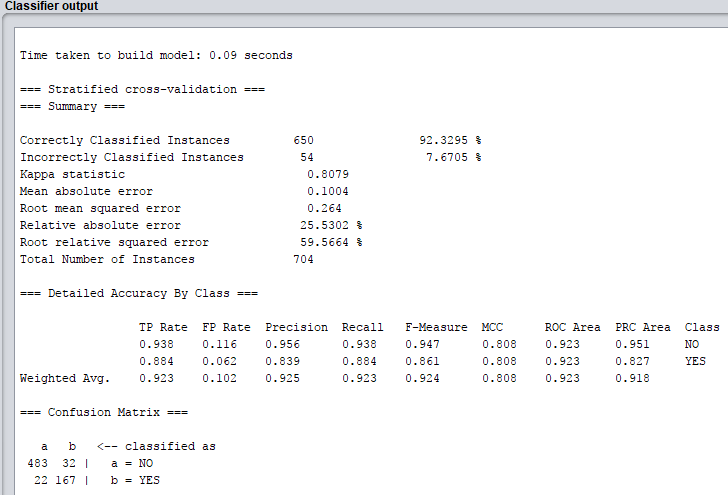
http://chem-eng.utoronto.ca/~datamining/dmc/zeror.htm

ZeroR is the simplest classification model and useful for determining a baseline performance for other classification methods. Without a baseline you do not know how well other algorithms are performing. Out of 704 instances in the dataset 515 were correctly classified (73%) and 189 were incorrectly classified (26%). With 73% accuracy this is useful for future classification methods to have a baseline performance, I can then evaluate if the other methods are useful in building a classification model. The ZeroR algorithm gives you a point of reference to which you can compare all other models that will be constructed.



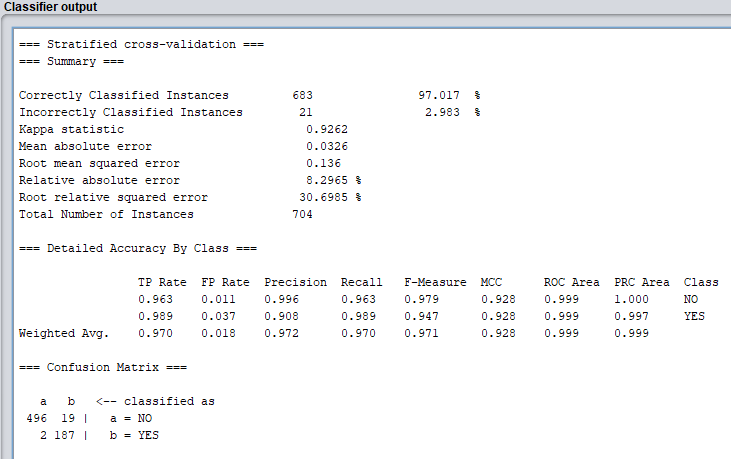
### Rules JRip

JRip algorithm produced a 92% correctly classified instances rate and 650 instances, with 7% incorrect, a total of 54 instances. The confusion matrix produced by the algorithm gave a total of aa + bb = 483 + 177 = 650 correctly classified instances, and ab + ba = 22 + 32 = 54 incorrectly classified instances. With a 92% accuracy using JRip this is a high accuracy but this does not take into account that the algorithm isn’t sensitive to class distribution, also with the fact that there is missing data in the dataset we don’t have access to all possible information. Taking this into account though 92% is still a very high accuracy. The kappa statistic of 0.8079 tells us that the classifier is doing better than chance and this high mark is a good indication that this method is performing well on the dataset.



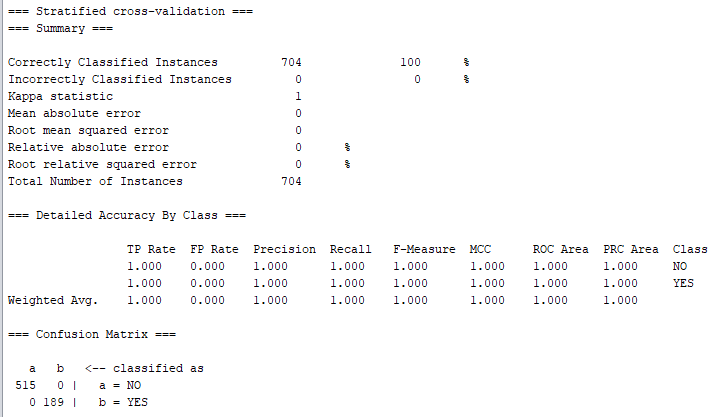
### Bayes NaiveBayes

NaiveBayes classification is used as an estimator class. Using numeric attributes it chooses which values based on analysis of the dataset. The algorithm uses the normal distribution to model numeric attributes, it can also handle numeric attributes using supervised discretization. This methods output produced one that is better than he JRip output. With 669 correctly classified instances, a 95% accuracy tells us that the algorithm performs well on the dataset and should be used in a classification model for this dataset. Only 4.9% inaccuracy and 35 incorrectly classified instances can be a good sign but taking into account the missing data a high accuracy could be a bad sign, how is the accuracy so high without imputing missing values. The true positives in NaiveBayes are very high at over 0.9, giving us a high count of correctly classified instances. The ROC Area value is also important, an optimal classifier will have ROC area values approaching 1, the ROC value produced by the NaiveBayes algorithm was a detailed accuracy by class of 0.999. This is an important value in terms of we know that the algorithm is making detailed and practical decisions on the class and is not random guessing, a ROC value of closer to 0.5 would mean that the classifier is random guessing.



### Functions SMO

The classifier SMO (Sequential Minimal Optimization) is implemented through globally replacing all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. The SMO algorithm also uses the RBF Kernel function and is used for solving the quadratic programming QP problem that arises during the training of support vector machines (SVM). Both the RBF Kernel and the SVM must be supplied by the user/machine in order for the SMO algorithm to run on a dataset. In our case the Autism-Adult dataset the SMO algorithm runs on the dataset and breaks the problem into a series of smallest possible sub-problems, which are then solved analytically. The times taken to run the algorithm was 0.06 seconds, the summary pointed out many things. Every instance (704) was correctly classified 100%, a = NO produced 515 instances and b = YES produced 189 instances. There was no error produced and the Kappa statistic was 1 meaning the classifier is performing well. Detailed accuracy by class report tells us that TP rate, Precision, Recall, F-measure MCC and ROC Area all performed at 1.00 and the false positive rate at 0.0.



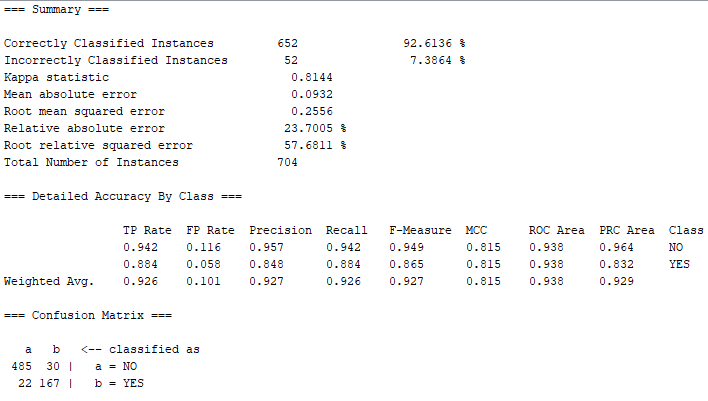
### Lazy IBk

Lazy IBk K-nearest neighbour’s classifier is used as a classifier algorithm that selects the appropriate value of K based on cross-validation. Options for running the algorithm include the weight neighbours by the inverse of their distance, the weight neighbours by 1 – their distance, the number of nearest neighbours used in classification, etc.

The optimal value of k I determined before was 15 as it gave the most accurate % classified instances and more instances were produced (676). It seems that from looking at the results the greater the value for K then the better and more accurate the output but this is not the case, I also tested for greater values of k such as 20 and 50, the accuracy of these models were less than the model of k=15.

### Trees J48

When running the J48 algorithm on the dataset the first number in the tree indicates the total number of instances reaching the leaf, the second number is the number (weight) of those instances that are misclassified. The number of correctly classified instances was 652 and 52 incorrect instances, 92% correct and 7% incorrect. Each of the detailed accuracy reports showed that TP Rate Precision and ROC area are all close to 1, this is a positive response and this classifier is one that should be using in a classification model.



The correctly and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified. The raw numbers are shown in the confusion matrix, with a and b representing the class labels. Here there were 100 instances, so the percentages and raw numbers add up, aa + bb = 59 + 12 = 71, ab + ba = 27 + 2 = 29.

The percentage of correctly classified instances is often called accuracy or sample accuracy. It has some disadvantages as a performance estimate (not chance corrected, not sensitive to class distribution), so you'll probably want to look at some of the other numbers. ROC Area, or area under the ROC curve, is my preferred measure.

Kappa is a chance-corrected measure of agreement between the classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. A value greater than 0 means that your classifier is doing better than chance (it really should be!).

The error rates are used for numeric prediction rather than classification. In numeric prediction, predictions aren't just right or wrong, the error has a magnitude, and these measures reflect that.

* **TP Rate**: rate of true positives (instances correctly classified as a given class)
* **FP Rate**: rate of false positives (instances falsely classified as a given class)
* **Precision**: proportion of instances that are truly of a class divided by the total instances classified as that class
* **Recall**: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)
* **F-Measure**: A combined measure for precision and recall calculated as 2 \* Precision \* Recall / (Precision + Recall)

As for the ROC area measurement, this is one of the most important values output by Weka. An "optimal" classifier will have ROC area values approaching 1, with 0.5 being comparable to "random guessing" (similar to a Kappa statistic of 0).

It should be noted that the "balance" of the data set needs to be taken into account when interpreting results. Unbalanced data sets in which a disproportionately large amount of instances belong to a certain class may lead to high accuracy rates even though the classifier may not necessarily be particularly good.